Proceedings of SALT 28: 409–432, 2018

Adjectival scales and three types of implicature*

Nicole Gotzner
Leibniz-ZAS, Humboldt-University

Stephanie Solt
Leibniz-ZAS

Anton Benz
Leibniz-ZAS

Abstract In this work, we explore the relationship between three different inferences triggered by gradable adjectives. In particular, we look at scalar implicature and two competing inferences occuring under negation - scale reversal (indirect scalar implicature) and a type of manner implicature called negative strengthening. In a series of experiments, we test a variety of adjectival scales and explore correlations between different inferences. Our results show that some scales are more likely to generate scalar implicature while others lean more towards generating negative strengthening. The extent to which scalar implicature and scale reversal correlate for the same scales, in turn, is lower than expected. We discuss our findings with respect to the mechanisms underlying the three types of inferences and factors accounting for differences across scales, with a focus on semantic distance, boundedness, the type of standard of comparison and adjectival extremeness.

Keywords: scalar implicature, scale reversal, negative strengthening, gradable adjectives, scale structure, negation

1 Introduction

The majority of work in formal and experimental pragmatics has focused on scalar implicature, while considerably less attention has been devoted to other kinds of pragmatic inferences and to the question of how different inferences might interact with one another. At the centre of our interest are different adjectival *Horn scales*. A Horn scale (Horn 1972; Levinson 1983) is a pair *strong/weak* of two expressions of comparable complexity that stand in a specific entailment relation: if *strong* occurs in

* We thank Richard Breheny, Eve Clark, Herbert Clark, Judith Degen, Napoleon Katsos, Manfred Krifka, Jacopo Romoli, Uli Sauerland, Chao Sun, Bob van Tiel as well as the audience of SALT 28 at MIT for insightful discussion. We are also grateful to Henry Salfner for assistance with the experiments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the Xprag.de Initiative (Grant Nr. BE 4348/4-1), a grant awarded to Stephanie Solt (grant Nr. SO 1157/1-2), and the Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411).

an upward entailing context $A(.)$, then $A(\textit{strong})$ semantically entails $A(\textit{weak})$, but not the other way around. We study the pragmatically licensed inferences that can be drawn from sentences $A(\textit{adj/neg adj})$ with a negated or non-negated adjectival scalar expression. As a concrete example, consider the scale *brilliant/intelligent* and the sentences in (1) and (2).

(1) a. She is intelligent.
b. She is brilliant.

(2) a. She is not intelligent.
b. She is not brilliant.

Semantic meaning allows for the inferences from (1b) *She is brilliant* to (1a) *She is intelligent*, and for the inverse inference from (2a) *She is not intelligent* to (2b) *She is not brilliant*. Explaining the pragmatically licensed inferences from utterances of positive sentences (1) to negative sentences (2), and vice versa, has been one of the core objectives of Gricean pragmatics (Grice 1975; Horn 1972, 1989; Levinson 2000). We focus on the three logically possible pragmatic inferences shown in (3).

(3) a. She is intelligent. $\rightsquigarrow$ She is not brilliant. (**SI**)
b. She is not brilliant. $\rightsquigarrow$ She is not intelligent. (**NegS**)
c. She is not brilliant. $\rightsquigarrow$ She is intelligent. (**SR**)

In (3a), the inferred *not brilliant* from an utterance of *intelligent* is called a *scalar implicature* (SI). In the Gricean tradition, it is explained as a consequence of the *maxim of quantity* (Grice 1975), which asks the speaker to be as informative as necessary. The inferences in (3b) and (3c) are drawn from the negated stronger scale mate. In (3b), *not brilliant* is strengthened to *not intelligent*. This inference is called *negative strengthening* (NegS, Horn 1989), and is traditionally explained as an inference from Grice's *maxim of manner*. The inference in (3c), on the other hand, is explained as a scalar implicature based on the reversed negative Horn scale *not brilliant/not intelligent*. Similarly, as *intelligent* implicates *not brilliant*, *not brilliant* implicates *not not intelligent*, hence, *intelligent*. This implicature involves a case of scale reversal (SR) since negation is a downward entailing operator. Chierchia 2004 calls this type of implicature *indirect scalar implicature* in order to stress the fact that the underlying mechanism is the same as for direct scalar implicature (see also Romoli 2012; Gotzner & Romoli 2018).

The inferences described above do not derive directly from semantic meaning but rather require the support of pragmatic principles. The resulting pragmatically enriched interpretations thus stand in competition with a purely semantic interpretation of the sentences in question. Furthermore, the three pragmatic inference types are

not completely independent of one another. In particular, negative strengthening and scale reversal are mutually exclusive, and, hence, cannot be valid in the same context. Additionally, since scalar implicature and scale reversal are assumed to originate from the same pragmatic principle, it is reasonable to expect that they will tend to arise in similar contexts and for the same pairs of items. To this point, there is now a considerable body of experimental literature demonstrating diversity in the rate at which pairs of scalar items (i.e. Horn scales) give rise to scalar implicatures (e.g., Doran, Baker, McNabb, Larson & Ward 2009; van Tiel, van Miltenburg, Zevakhina & Geurts 2016). In earlier work (Gotzner, Solt & Benz 2018) we have demonstrated a similar pattern of diversity in rates of negative strengthening, and furthermore an anti-correlation between scalar implicature and negative strengthening rates. However, it was unclear to what extent this was a task-related effect or rather an indication of a deeper relationship between the two types of inferences. Furthermore, to date there has been no investigation of how scale reversal implicatures pattern with regards to 'scalar diversity'. In this paper, we present the results of a series of experiments investigating the three above-described pragmatic inferences among a broad variety of adjectival Horn scales. Our goal is to provide a systematic picture of the relationships between the three inference types, as well as the factors that contribute to the presence or absence of each. In doing so, we seek to shed light on the mechanisms on which these different pragmatic inferences are based.

This paper is structured as follows. In Section 2, the relevant neo–Gricean accounts of scalar implicature, scale reversal under negation, and negative strengthening are introduced. In Section 3, previous studies on scalar diversity and implicatures of adjectival scales are reviewed. Section 4 presents a series of experiments testing the relationship between scalar implicature, negative strengthening and scale reversal across adjectival scales. Finally, we discuss the relationship of the three types of implicature and factors that account for variability across scales.

2 Three types of implicature

Horn 1972 approaches the topic of scalar implicature from the perspective of Aristotle's square of opposition, as depicted in Fig 1. The square shows the logical relations between four sentences situated in the A, E, I, and O corners. The Aristotelian example for A, E, I, and O are the quantified sentences with *all* in the A and *some* in the I corner. The sentences in opposite corners A–O and I–E are contradictories, meaning that exactly one of the two sentences is true. The contraries in the A and E corners are also mutually exclusive (i.e., *A and E* is a contradiction), but it is conceivable that neither A nor E is true. It was Horn's insight that the sub-contrary relation between the E and O corners can be explained by pragmatic inferences based on Grice's maxim of quantity. If both the A sentence (*all*) and the I sentence (*some*)

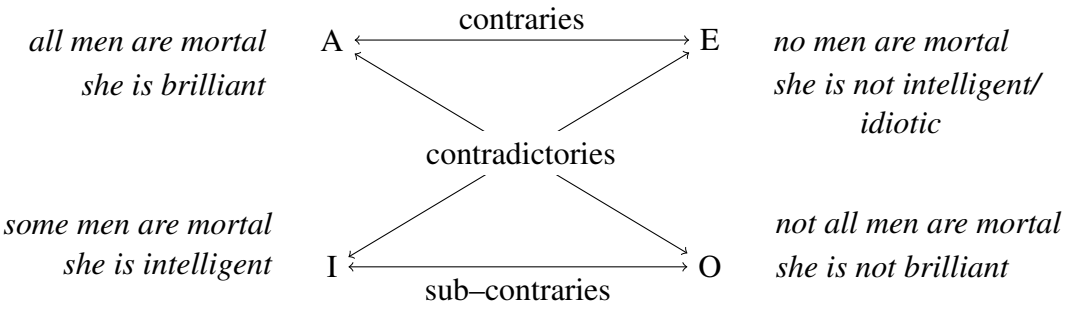

Figure 1 Aristotle's square of opposition for *all/some* and *brilliant/intelligent*

can be truthfully asserted, a speaker following the maxim of quantity must prefer the A sentence. Therefore, a speaker who uses the I sentence will only do so if the A sentence is false. Hence, the I sentence *conversationally implicates* the O sentence.

Horn also saw that the negation on the right side of the square gives rise to the reverse implicature $O \rightsquigarrow I$. This can be seen if we write *none* as *not some*. Negation creates a downward entailing context which reverses the positive scale *all/some*. As in the case of A and I, the maxim of quantity then implicates that a speaker using O (*not all*) implicates that I (*not not some* = *some*). Thus this visualization of the meaning relations between sentences captures the dual character of scalar implicature (SI) and scale reversal (SR).

It might be tempting to take Aristotle's square of opposition as a template to be applied to all kinds of Horn scales. However, it is particularly important in the context of adjectival scales that the meaning relations of the square of opposition do not generalise. As an example, let us again consider the Horn scale *brilliant/intelligent* with the positive sentences *She is brilliant/intelligent* from (1) and the negative ones *She is not brilliant/intelligent* from (2). In the square of opposition, the strong scale element *brilliant* sits in the A corner, and the weak *intelligent* in the I corner. As for *all/some*, the negated counterparts sit in opposite corners: *not intelligent* in E and *not brilliant* in O. The contradictory and the sub–contrary relations hold as in the *all/some*-case. The relation that does not generalise is the *contrary* relation: the pragmatically relevant contrary to *brilliant* is not *not intelligent* but rather an antonym such as *idiotic*. However, if the contrary *idiotic* is put in the E corner, then there is a semantic gap between it and *intelligent* in the I corner, i.e. the *contradictory* relation between I and E no longer holds.

In cases such as these, the crucial meaning relations are better visualized not via the square of opposition, but instead by relating adjectival scale-mates to their corresponding territories on the underlying *measurement scale*. In the case of *bril-*

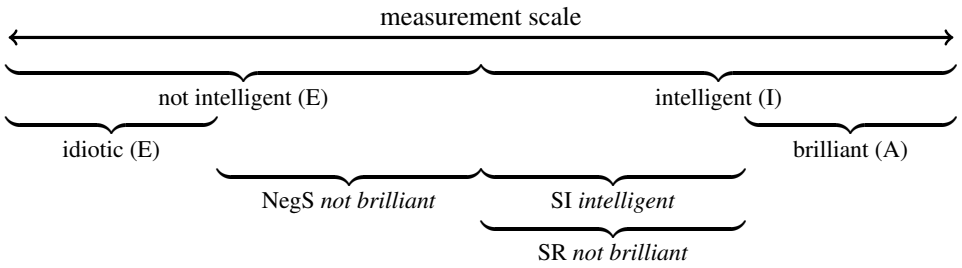

Figure 2 Meaning extensions on the measurement scale for Horn scale *brilliant/intelligent*. Corners of the square of opposition are marked by (A), (E), and (I), with two alternatives for E. The O corner is the logical *not brilliant*. Pragmatically strengthened meanings are in *italics*.

liant/intelligent, the measurement scale reaches from extreme states of intelligence to extreme stupidity, as shown in Fig. 2, which depicts the semantic meanings of the strong (A) and weak (I) scale-mates as well as the two alternatives for the E corner. Crucially, there is a scalar gap between the I term (*intelligent*) and the stronger choice for the E term (*idiotic*), which is not lexicalized by any simple (i.e. non-negated) expression. Horn 1989 proposed that it is this very type of situation that gives rise to negative strengthening (NegS) – an implicature by which the negative O sentence (*not brilliant*) is strengthened such that it fills this gap. In the neo–Gricean tradition, this strengthening implicature is explained as a consequence of Grice's maxim of manner: As a speaker prefers the less marked *intelligent* over the phrase *not brilliant*, if *intelligent* can be truthfully asserted, it follows that an occurrence of *not brilliant* must imply that *intelligent* is not applicable.[1]

As seen in Fig. 2, the results of negative strengthening (NegS) and scale reversal (SR) implicatures are mutually incompatible interpretations for *not strong*. The former arises via a manner implicature that the *weak* term (i.e. *intelligent*) does not obtain, resulting in a strengthened meaning 'less than *weak*'. The latter is derived as a scalar implicature assuming the contradictory *not weak* in the E corner, resulting in the meaning 'not *strong* but *weak (=not not weak)*; on the latter interpretation, the negated *strong* term covers the same scalar range as the non-negated *weak* term enriched via scalar implicature (SI).

1 Alternative accounts of negative strengthening are Blutner's 2004 optimality theoretical framework and Krifka's 2007 partial blocking framework, which is based on iterative application of Levinson's M–principle (Levinson 2000). Horn 2017 explains negative strengthening of *not adj* by considering the square with *adj* in A, *not adj* in O, and the morphologically negated adjective *neg-adj* in E position; that is, he considers the square without the weak scale–mate and, therefore, cannot make inference about the weak term. However, none of these models make the fine–grained predictions that are necessary for explaining scalar diversity.

Importantly, conceptualizing the meaning relationships among adjectival Horn scale-mates with reference to the underlying measurement scale suggests the possibility that properties of this scale may have an influence on which of the above inferences are generated. As is well known, adjectival scales differ in their structures, in particular with respect to the presence or absence of scalar endpoints (Kennedy & McNally 2005). Adjectives themselves may denote scalar endpoints (*clean*) or the non-endpoint portion of a scale (*dirty*), or may have purely contextual standards (*big*); they may denote extreme scalar values (*gigantic*) or values close to the origin point (*damp*); and they may enforce a higher standard of precision (*spotless*) or a more relaxed one (*cleanish*). This rich variation in scale structures and corresponding adjective meanings makes the adjectival domain an ideal one for exploring the interplay of the three inference types discussed here.

3 Previous experiments

From a theoretical perspective, scalar implicature is predicted to be triggered for any pair of weak and strong scale-mates. However, several experimental studies have demonstrated that the extent to which participants compute a scalar implicature varies considerably across different Horn scales (Doran, Baker, McNabb, Larson & Ward 2009; Doran, Ward, McNabb, Larson & Baker 2012; Beltrama & Xiang 2013; van Tiel, van Miltenburg, Zevakhina & Geurts 2016; Simons & Warren 2018; Benz, Bombi & Gotzner 2018; Gotzner, Solt & Benz 2018). Van Tiel et al. (2016) investigated scalar implicature rates among 43 *weak/strong* pairs using a task in which participants were presented with an utterance by a speaker including a weak scalar term (e.g. 'John says: she is intelligent') and were asked to judge whether the negation of a stronger scale-mate obtained (e.g. 'Would you conclude from this that, according to John, she is not brilliant?'). They found considerable variability in the rates of implicatures measured in this way. They further investigated the factors accounting for this so-called scalar diversity and found that boundedness and semantic distance explained a modest part of the variability. For example, participants were more likely to derive a scalar implicature for the bounded *some/all* scale than the unbounded *warm/hot* scale. Semantic distance was measured by a participant rating of the relative strength between the statements involving the weaker and stronger scale-mates (e.g., the pair *difficult/impossible* had a high distance rating). However, most of the variance in the van Tiel et al. study remained unexplained, and there was also considerable overlap between the factors boundedness and grammatical category, e.g. most unbounded scales were adjectival ones (see Gotzner et al. 2018 for further discussion).

Gradable adjectives have also been shown to differ with respect to the inferences triggered under negation. In particular, Leffel, Cremers, Gotzner & Romoli

forthcoming found that minimum standard adjectives yielded an inference to the positive form in the 'not very' construction (*John was not very late* $\rightsquigarrow$ *John was late*) while relative ones like *tall* were negatively strengthened, indicating a role of scale structure in implicature computation. Further, an investigation by Ruytenbeek, Verheyen & Spector 2017 tested the role of (evaluative) polarity and morphological complexity in negative strengthening, looking at pairs of antonyms. The authors found that positive terms like *happy* were more likely to be negatively strengthened than their negative antonyms and there was also a difference with respect to morphological and non-morphological pairs, for example *unhappy* and *sad* (see also Tessler & Franke 2018 for a computational model). This work is in line with the suggestion by Horn 1989 that politeness considerations are relevant to negative strengthening (it is also in keeping with Krifka's and Blutner's idea that blocking and conventionalization play a role). However, Leffel et al. forthcoming provide examples showing that politeness is orthogonal to the effect of scale structure.

Previous work by two of the present authors (Benz et al. 2018) investigated the relationship between scalar implicature and negative strengthening in the paradigm by van Tiel et al. 2016. As noted above, in the van Tiel et al. study, participants were presented with statements such as *she is intelligent* and were asked whether they would infer that *she is not brilliant*. We hypothesized that the use of negated adjectives in the conclusion sentence brings into play negative strengthening. That is participants may respond NO to the conclusion because they interpret *not brilliant* in a strengthened manner as 'not intelligent'; on this strengthened interpretation, the conclusion sentence is actually incompatible with the antecedent sentence. In a corresponding experiment that assessed endorsement of negative strengthening (see Table 2), we found that scalar implicature was anti-correlated with degree of negative strengthening of the stronger scale-mate. This was taken as potential evidence that the apparent 'scalar diversity' found in van Tiel et al.'s study was in part caused by the masking of scalar implicature by negative strengthening.

In Gotzner et al. 2018, we expanded the investigation of scalar implicature and negative strengthening to a broader set of adjectival pairs. Adjectives represent a rich ground for exploring this topic because they are open class words, allowing ample stimulus items to be created in which factors previously shown to play a role in predicting implicature rates are systematically varied. We created a set of 70 adjective pairs which varied in several dimensions of scale structure, including boundedness, extremeness, polarity and distance between scale-mates. These were used as the basis for two inference tasks: a scalar implicature task using the methodology of van Tiel et al. 2016 and a negative strengthening task as conducted by Benz et al. 2018. Again, we found an anti-correlation between endorsement rates for the two types of inferences. Furthermore, it was found that they share many of the same predictors: endorsements of scalar implicature were higher for upper-bounded scales

and more distant scale-mates and higher for negative vs. positive scales (defined in the dimensional sense). In turn, scalar implicature rates were lower when the strong scale-mate was an extreme adjective like *stunning* (based on diagnostics by Morzycki 2012). Negative strengthening rates, on the other hand, were higher for extreme adjectives and lower for more distant scale-mates (see Gotzner et al. 2018 and the Appendix for a full description of all predictors).

In summary, previous experimental research has demonstrated diversity in the types and rates of pragmatic inferences that arise for different Horn scales, as well as patterns of (anti-)correlation between the difference inference types (particularly between scalar implicature and negative strengthening). It has also been demonstrated that factors related to the underlying scalar semantics of gradable adjectives play a role in determining to what extent such items participate in various types of pragmatic inferencing. However, the existing studies do not allow it to be conclusively determined whether the observed anti-correlation between scalar implicature and negative strengthening is a purely task-related effect (the masking of scalar implicature by negative strengthening), or whether it represents a deeper connection between the two phenomena. Nor has it been investigated to what extent the endorsement of scale reversal implicatures might be correlated with those for the other inference types. Clarifying these points will provide a deeper understanding of the systematic relationships between the different inferences that arise from pairs of strong/weak scale-mates, and potentially provide insight into the mechanisms underlying these phenomena.

4 Current Experiments

4.1 Goals of current experiments

In Gotzner et al. 2018 we reported the results of two experiments in which we tested scalar implicature with the original task of van Tiel et al. 2016 as well as well as negative strengthening for 70 adjectival Horn scales. In this paper, we present two new tasks and compare them to the previously reported ones.

The first goal of our study is to test whether the anti-correlation between scalar implicature and negative strengthening that was found by Gotzner et al. 2018 holds more generally. Specifically, we seek to rule out the possibility – left open by our previous study – that the observed anti-correlation was primarily a task-related effect. The alternate hypothesis that we explore is that some scales are inherently more likely to give rise to scalar implicature, while others tend to give rise to negative strengthening.

To test this possibility, we devised a modified scalar implicature task which blocks negative strengthening in the conclusion sentence. That is, rather than

simply presenting participants with the potential scalar implicature (the negated stronger scale mate), we presented them with the enriched meaning, i.e. the literal meaning taken together with the scalar implicature. For example, participants judged whether the statement *He is attractive* suggests that according to the speaker *He is attractive but not stunning* (by contrast, the original task presented the conclusion *He is not stunning*). In the modified task, negative strengthening is not an available reading anymore, because mentioning the weaker term sets a lower bound on the interpretation, so that the statement cannot mean 'rather unattractive'.

A second goal of our current experiments was to investigate how scale reversal implicature factors into the picture. With the combination of (direct) scalar implicature, negative strengthening and scale reversal we aim to better be able to classify different triggers and the inferences they give rise to. While not explicitly stated in the theoretical literature, it is reasonable to assume that scales that are likely to generate scalar implicature are also likely to generate scale reversal implicature under negation, since the two inferences are thought to be based on the same mechanism. On the other hand, since scale reversal and negative strengthening are contradictory, participants should either derive the former inference or the latter one. Hence, we predict to find an anti-correlation between scalar implicature and negative strengthening as well as between negative strengthening and scale reversal. Further, there should be a positive correlation between scalar implicature and scale reversal.

4.2 Methods

4.2.1 Participants

Participants with US IP addresses were recruited on Amazon's Mechanical Turk platform and were further screened for native language. In total, 80 native English speakers (mean age: 35.3, 35 female, 45 male) took part in the study. They were paid 1 dollar in compensation.

4.2.2 Materials

Our materials were based on a set of 70 adjective pairs with weak and strong scale-mates, the same pairs that were used in the experiments reported in Gotzner et al. 2018.[2]We took all adjective pairs from the van Tiel et al. study (32) and added a further set of 38 adjective pairs to balance factors related to the scale structure of the adjectives. In particular, we added further absolute gradable adjectives (minimum standard and maximum standard), as well as more pairs where the stronger scale-

2 The original list contained 71 pairs, but the pair *content/unhappy* was excluded from further analyses on the basis of diagnostics showing that the two terms are not on the same scale.

Mary says:

He is intelligent.

Would you conclude from this that, according to Mary, he is intelligent but not brilliant?

□Yes □No

Figure 3 Sample item of the modified scalar implicature task.

mate is non-extreme. A complete list of all materials can be found in the Appendix of Gotzner et al. 2018 and in a repository on OSF (https://osf.io/muahf/).

These 70 adjective scales were embedded in two separate tasks administered to 40 participants each. In the current paper, we compare ratings across four different tasks, the two which were reported in Gotzner et al. 2018 and two new tasks. Table 1 presents an overview of the tested inferences in Gotzner et al. 2018 and the two new tasks. The first new task was a modified version of the scalar implicature task used in van Tiel et al. 2016 and Gotzner et al. 2018. Specifically, we added the weak term in the conclusion sentence so that negative strengthening is blocked. Essentially, the conclusion sentence now presents the enriched meaning, i.e. the literal meaning taken together with the scalar implicature. A sample stimulus item for this task is shown in Figure 3.

Task	Statement	Candidate Inference
SI_original (GSB)	Mary says: "He is intelligent"	He is not brilliant
SI_mod (current)	Mary says: "He is intelligent"	He is intelligent but not brilliant
NegS (GSB)	Mary says: "He is not brilliant"	He is not intelligent
SR (current)	Mary says: "He is not brilliant"	He not brilliant but he is intelligent

Table 1 Overview of statements and candidate inferences: SI_original and NegS were reported in Gotzner et al. (2018) (GSB). SI_mod and SR are the two new inference tasks we report here (current).

Our second new task was designed to measure scale reversal implicatures across different triggers. We devised a version of the conclusion sentence that corresponded to the scalar implicature task in that the literal meaning taken together with the scale reversal implicature was presented in the conclusion sentence.[3] Examples of the stimulus sentence and the conclusion sentence for this task as well as the negative strengthening task from Gotzner et al. 2018 are shown in Table 1. Note that

3 We also ran a second version of this task where we only presented the statement with the weaker term in the conclusion sentence, see the Appendix.

participants only saw one of the conclusion sentences.

Finally, by means further experimental tasks as well as manual annotation, each of the 70 adjective pairs was profiled on a range of dimensions hypothesized to play a role in the frequency at which implicatures of different sorts are drawn. Those dimensions were: boundedness (whether the stronger scale-mate denotes an endpoint on the underlying measurement scale); standard type (whether the weaker scale-mate invokes a minimum standard, maximum standard or contextually dependent relative standard); extremeness (whether the stronger scale-mate is an extreme adjective in the sense of Morzycki 2012); the relative frequency of the weak vs. strong scale-mates; polarity (positive or negative); the perceived semantic distance between the weak and strong scale-mates; and the politeness of the weak term, the strong term, and the negated strong term. The experimental tasks and annotation procedures used in these classifications are described in detail in Gotzner et al. 2018. Table 4 in the Appendix presents an overview of all tasks and Table 5 details the annotation of additional factors.

4.3 Results

The mean endorsement rate in the modified scalar implicature task was 66.2%, higher than that of the original task reported in Gotzner et al. 2018 (39.2%). The mean endorsement rate for the scale reversal task was 54.7%, and that for the negative strengthening task (from Gotzner et al. 2018) was 60.0%. However, in the case of all of these implicature types, there was a high degree of variability in the endorsement rates across the pairs tested. For scalar implicature on the modified task reported in the present paper, endorsement rates varied from 41% (for the pair *dirty/filthy*) to 90% (for *cheap/free*); for scale reversal, the range was 29% (*sickish/sick*) to 82% (*big/enormous*); and for negative strengthening, it was 29% (*sick/terminally ill*) to 90% (*thin/skinny*).

Table 1 presents sample rates for all different tasks together with an annotation of factors related to scale structure.

To assess the interplay of the three inference types, we computed correlation tests between all different tasks. Figure 4 displays the correlations between all tasks based on Pearson's correlation test. We also computed Kendall's tau to compare the pairwise rankings across tasks. There was a positive correlation between the original scalar implicature task and the modified task (r_τ = .63, p <.0001). As previously observed, there was a negative correlation between the scalar implicature and the negative strengthening task (modified task: r_τ = -.47, p <.0001; original task: r_τ = -.49, p <.0001). Further, the negative strengthening task was negatively correlated with the scale reversal task (r_τ = -.38, p <.0001). Finally, there was a positive correlation between the scalar implicature and scale reversal task (modified task: r_τ

Weak/strong term	Scale structure	SI	SI_mod	NegS	SR
cheap/free	bounded rel neg non-extreme	0.76	0.90	0.41	0.54
possible/certain	bounded min pos non-extreme	0.58	0.87	0.3	0.78
clean/spotless	bounded max neg extreme	0.27	0.51	0.75	0.56
wet/soaked	unbounded min pos extreme	0.24	0.54	0.44	0.77
large/gigantic	unbounded rel pos extreme	0.22	0.64	0.74	0.69
scared/petrified	unbounded rel neg extreme	0.14	0.57	0.75	0.43

Table 2 Example scales and their respective endorsement rates in the original (SI) and modified scalar implicature (SI_mod), negative strengthening (NegS) and scale reversal (SR) task

= .25, p = .01; original task: r_τ = .32, p <.0001).

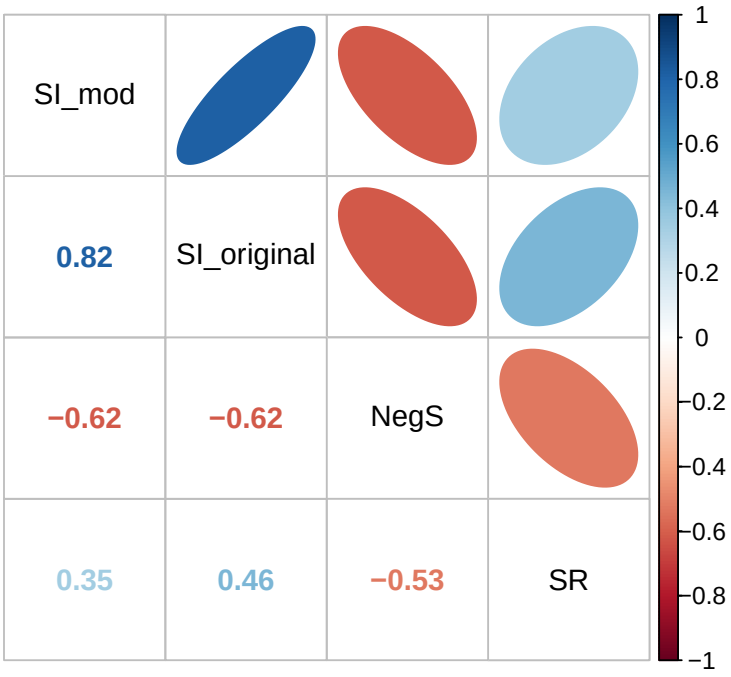

Figure 4 Pearson's correlations between the modified scalar implicature task (SI_mod), the original one (SI_original), negative strengthening (NegS) and scale reversal (SR). Positive correlations are shown in blue and negative ones in red.

We were also interested in what predictors account for variability across triggers. The original models reported in Gotzner et al. 2018 for scalar implicature found

effects of boundedness, the standard invoked by the weaker term (e.g., maximum standard *clean* vs. relative standard *large*), polarity, semantic distance and extremeness. For negative strengthening main predictors were extremeness, distance and the type of standard by the weaker term, with opposite direction of the effects (see Table 8 in the Appendix). The predictors in the modified scalar implicature task were the same as in the original task except that polarity did not have a significant effect. That is, scalar implicature rates were higher for upper-bounded scales ($p < .01$) and higher for more distant scale-mates ($p < 0.001$) while endorsements were lower for extreme stronger terms ($p < 0.001$) and weak terms denoting a scalar maximum ($p < 0.05$), for example *clean* vs. the relative weak term *annoyed*.

In the scale reversal task, more distant scale-mates also yielded higher endorsement rates ($p < 0.001$) as well as pairs with a a higher relative frequency of the weak relative to the strong term ($p < 0.05$). Extreme stronger terms had marginally lower endorsement rates than non-extreme ones ($p = 0.054$). In Tables 6, 7 and 8 in the Appendix, we detail the predictors for the modified scalar implicature task, the scale reversal task and the negative strengthening task reported in Gotzner et al. 2018.

Given the anti-correlation that was found between scalar implicature and negative strengthening, we would like to capture which pairs of scale-mates lean more towards triggering one of these two sorts of inferences versus the other. Therefore, we created a unified vector for each pair. We first normalized each vector with a min/max function $\left(\frac{x-Min(x)}{Max(x)-Min(x)}\right)$. Then, we created a unified vector for scalar implicature and negative strengthening with the formula $\frac{SI}{SI+NegS}$. The same was done for the ratio between negative strengthening and scale reversal, which were likewise anti-correlated.

We fit linear regression models with all original factors outlined in Gotzner et al. 2018 for both the unified measure of SI and NegS as well as a unified measure of NegS and SR. The first unified measure showed that pairs with maximum standard and extreme terms were leaning more towards triggering NegS ($p < 0.05$ and 0.01) while upper-bounded and more distant scale mates were more likely to trigger SI ($p < 0.05$ and $p < 0.001$). The second unified measure revealed that maximum standard and extreme terms were more likely to trigger NegS than SR ($p < 0.05$ and 0.01) while upper-bounded and more distant scale mates were more likely to trigger SR ($p < 0.05$ and $p < 0.001$).

Hence, variability in both measures was mainly driven by the same predictors: upper boundedness, the standard invoked by the weaker term[4], semantic distance and extremeness. The detailed models are presented in Tables 9 and 10 in the Appendix.

4 We also ran both models with minimum standard weaker terms as the reference level and there was a significant difference between minimum and maximum standard adjectives ($p < .05$) but not difference between relative and minimum standard adjectives.

predictor	SI_mod	NegS	SR
boundedness	yes	–	–
standard	yes	yes	–
distance	yes	yes	yes
extremeness	yes	yes	(yes)
frequency	–	–	yes

Table 3 Predictors of the modified scalar implicature (SI_mod), negative strengthening (NegS), and scale reversal (SR) task.

Table 3 presents an overview of the predictors for the individual models for the modified SI, NegS, and SR task.

5 Discussion

5.1 Summary of findings

In a series of experiments, we investigated the extent to which 70 adjective pairs with weaker and stronger scale-mates give rise to three types of pragmatic inferences: scalar implicature (SI), negative strengthening (NegS) and scale reversal (SR). We found (i) variability in inference rates across all three types of implicature and (ii) that these three inferences are related in a specific way.

Our investigation had two goals: first, we wanted to test whether there is a deeper connection between SI and NegS; and, second, we wanted to investigate the relationship of these two inferences with SR. We devised a modified SI task that blocks negative strengthening as a task-related effect (by mentioning the weaker scale-mate again in the conclusion sentence) and we again found a strong anti-correlation between SI and NegS rates. Importantly, we also found the same predictors accounting for variability across scales in the original and the modified task.[5] As predicted, we observed that NegS and SR are anti-correlated and our experiments also showed a weaker positive correlation between SI and SR (see Table 4).

Overall, our results indicate that some pairs of scale-mates are more likely to generate SI while others are leaning more towards NegS. We demonstrated this by using combined measures of scalar SI and NegS as well as NegS and SR. The combined measures showed that factors related to vagueness and scale structure are crucial in determining where along the continuum a pair of expressions falls, in particular upper boundedness, the type of standard invoked by the weaker term, semantic distance and adjectival extremeness.

5 Polarity, however, was not a significant predictor in the modified task anymore, potentially due to the changed wording with *but* in the conclusion sentence.

5.2 Interplay of scalar implicature, negative strengthening and scale reversal

As mentioned in Section 4.1, it is generally assumed that SI and SR are both scalar quantity implicatures. This leads to the natural expectation that scales with high SI-rates have also high SR-rates, and vice versa. As SR and NegS are logically incompatible, this should entail an anti–correlation between SI- and NegS-rates. As Table 4 shows the predicted correlations are borne out, however, the positive correlation between SI and SR is weaker than expected. It, therefore, seems that the anti–correlation between SI and NegS cannot simply be explained as an indirect effect of the positive correlation between SI and SR, and the logical incompatibility between SR and NegS. If we consider the predictors of the different implicature types in Table 3, we find a further argument against this indirect explanation. SR and SI share one predictor only, namely semantic distance, whereas, SI and NegS share distance, standard invoked by the weaker term, and extremeness. This suggests that there is a more direct explanation of the anti–correlation between SI and NegS, which relates to their underlying common predictors.

In Section 2, we saw that the standard square of opposition as defined for *all/some* does not generalise to adjectival scales as the pragmatically relevant contrary to the strong adjective is an antonymic phrase which is, in general, not the contradictory of the weak scalar adjective. Our results show that the *semantic distance* between weak and strong scale-mates has an effect on all three types of implicature: as semantic distance increases the SI– and SR–rate increases and the NegS–rate decreases. We may think of semantic distance as the distance on a measurement scale between the lower bounds of the intervals defined by the weaker scale–mate W and the stronger one S. As the distance between the lower bounds increases, the more likely it becomes that the speaker means by saying W that S is not the case, and, hence, it is more likely that the hearer derives a scalar implicature. Negative strengthening (NegS) is explained as a *blocking* phenomenon (Horn 1989; Levinson 2000; Blutner 2004; Krifka 2007). That is, the existence of the unmarked expression *blocks* parts of the semantic meaning of the marked expression. If the distance between W and S widens, W has to block a larger interval on the underlying measurement scale, and it may therefore become less probable that W succeeds in doing this. As a result, NegS rates will decrease with increasing semantic distance, as depicted in Figure 5. If the meaning of W is not blocked, the possibility for scale reversal implicature opens up. Hence, as semantic distance increases SR-rates should increase, too. There is, however, an asymmetry between NegS and SR: NegS strengthens the meaning of *not S* to an interval on the measurement scale for which there is no salient lexical expression, whereas SR implicature strengthens it to an interval that is already covered by the weak term.

Another consequence of our considerations of the square of opposition is that

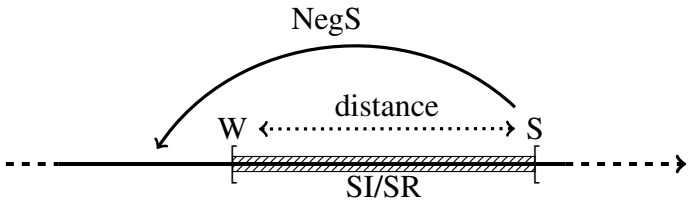

Figure 5 Semantic distance and negative strengthening. W and S are the lower bounds of the weak and strong scalar term, respectively, on an underlying measurement scale. As the distance increases, the more likely it is that an utterance of W implicates that S is excluded (SI), and the less likely that the negation of S jumps over W into the region below W (NegS).

NegS can only occur if there is a gap between the meaning of the weak scale–mate and the contrary of the stronger one (see especially Horn 2017 for a recent proposal that incorporates this idea). However, this is only the case if the pragmatically relevant contrary is an antonym of the strong scale–mate that is different from the logical contradictory of the weaker one. For SR–implicature, there is no obvious reason why the existence of a gap between weak scale–mate and the contrary of the strong one should increase or decrease SR–rates. This may explain why the difference between minimum, maximum and relative standard adjectives is, indeed, not significant for SR–rates.

Our investigation showed that scales in which the stronger term denotes a scalar endpoint tend towards SI implicatures, whereas those in which the stronger term has a vague interpretation are more likely to give rise to NegS.[6] We now turn to discussing the relevance of vagueness and extremeness in inference computation.

In a previous study by Leffel et al. forthcoming, minimum standard adjectives but not relative adjectives triggered an inference to the positive form in the 'not very' construction (e.g., *John was not very late* implicated that John was late). The authors offered an account for the role of vagueness in implicature computation that incorporates a borderline constraint. For example, when there are no heights that clearly count as *tall* and *tall but not very tall* at the same time, no inference negating the stronger term is derived. So, essentially the assumed role of vagueness is to determine which terms serve as good alternatives. This account explains why hearers are less likely to draw scalar inferences with vague terms. It is also in line with

6 The exception to this generalization is the case in which the weaker term itself is a maximum standard adjective and the stronger term denotes this endpoint interpreted at a higher level of precision. Thus apparently scales based on manipulation of precision level behave differently. We refer the reader to Gotzner et al. 2018 for further discussion.

the positive effect of semantic distance on scalar implicature rates in our SI–tasks: hearers may be hesitant to endorse a scalar implicature when the pair of scale-mates does not differ enough in perceived strength since the likelihood may be greater that both the weak and strong term apply. If, on the other hand, one of the two terms is bounded, the weak and the strong term are clearly distinguishable, even when the semantic distance between expressions is low. This would explain why bounded scales have higher SI–rates.

The effect of extremeness on implicature computation is two-fold. On the one hand, we found that extreme adjectives were less likely to trigger SI and SR. This could be could due to fact that the weak and extreme strong terms are used in different contexts, in line with Morzycki's 2012 view that extreme adjectives signal that the degree lies outside of the contextual range. Therefore, an extreme stronger term may not come into mind when the speaker uses W and no scalar implicature is derived. On the other hand, we found that extreme adjectives were good candidates for NegS. When uttering a statement with a negated extreme term like *John is not stunning*, a speaker makes a very underinformative statement and this could give the hearer a cue to reason about why the speaker did not want to make a commitment. Thus, speakers may use negated extreme terms to invite a pragmatic strengthening while leaving the literal meaning rather weak.

Overall, the strong anti-correlation we find between SI and NegS suggests that (i) the two inferences might be more related than previously-thought or that (ii) something else needs to be said about alternatives and Horn scales (for example see Horn 2017 for a proposal that alternatives of different complexity may function as scale-mates). Further, NegS and SR may be similar types of inferences but what determines which inference is drawn is the gap between expressions.

6 Conclusions

In conclusion, our results indicate that there is a closer connection between negative strengthening and scalar implicature while scalar implicature and scale reversal could be less directly related. It is hence an open question whether distinct pragmatic principles need to be postulated to explain these inferences. Our investigation suggests that crucial factors determining which inference is derived relate to vagueness, that is, the existence of a gap, boundedness and the type of standard of comparison; in addition to adjectival extremeness.

References

Beltrama, Andrea & Ming Xiang. 2013. Is excellent better than good? Adjective scales and scalar implicatures. In *Sinn und Bedeutung 17*, 81–98.

Benz, Anton, Carla Bombi & Nicole Gotzner. 2018. Scalar diversity and negative strengthening. In Uli Sauerland & Stephanie Solt (eds.), *Sinn und Bedeutung 22*, vol. 1 ZAS Papers in Linguistics 60, 191–204. Berlin: ZAS.

Blutner, Reinhard. 2004. Pragmatics and the lexicon. In Lawrence Horn & Gregory Ward (eds.), *The Handbook of Pragmatics*, 488–514. Oxford: Blackwell Publishing.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena, and the syntax / pragmatics interface. In Adriana Belletti (ed.), *Structures and Beyond*, 39–103. Oxford: Oxford University Press.

Doran, Ryan, Rachel E. Baker, Yaron McNabb, Meredith Larson & Gregory Ward. 2009. On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics* 1. 211–248. doi:https://doi.org/10.1163/187730909X12538045489854.

Doran, Ryan, Gregory Ward, Yaron McNabb, Meredith Larson & Rachel E. Baker. 2012. A novel paradigm for distinguishing between what is said and what is implicated. *Language* 88. 124–154. doi:10.1353/lan.2012.0008.

Gotzner, Nicole & Jacopo Romoli. 2018. The scalar inferences of strong scalar terms under negative quantifiers and constraints on the theory of alternatives. *Journal of Semantics* 35(1). 95–126. doi:https://doi.org/10.1093/jos/ffx016.

Gotzner, Nicole, Stephanie Solt & Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9. doi:10.3389/fpsyg.2018.01659.

Grice, Herbert Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and Semantics*, vol. 3, 41–58. New York: Academic Press.

Horn, Laurence R. 1972. On the Semantic Properties of the Logical Operators in English.

Horn, Laurence R. 1989. *A Natural History of Negation*. Chicago: University of Chicago Press.

Horn, Laurence R. 2017. Lie-toe-tease: double negatives and unexcluded middles. *Philosophical Studies* 174. 79–103. doi:10.1007/s11098-015-0509-y.

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81. 345–381. doi:10.1353/lan.2005.0071.

Krifka, Manfred. 2007. Negated antonyms: Creating and filling the gap. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and Implicature in Compositional Semantics*, 163–177. London: Palgrave Macmillan.

Leffel, Tim, Alexandre Cremers, Nicole Gotzner & Jacopo Romoli. forthcoming. Vagueness and the derivation of structural implicatures. *Journal of Semantics* .

Levinson, Stephen C. 1983. *Pragmatics. Cambridge Text Book in Linguistics*. Cambridge: Cambridge University Press.

Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicatures*. Cambridge, MA: MIT Press.

Morzycki, Marcin. 2012. Adjectival extremeness: Degree modification and contextually restricted scales. *Natural Language & Linguistic Theory* 30(2). 567–609.

Romoli, Jacopo. 2012. *Soft but strong. Neg-raising, soft triggers, and exhaustification*. Boston, Massachusetts: Harvard University PhD dissertation.

Ruytenbeek, Nicolas, Steven Verheyen & Benjamin Spector. 2017. Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: A Journal of General Linguistics* 2. 1–27. doi:10.5334/gjgl.151.

Simons, Mandy & Tessa Warren. 2018. A closer look at strengthened readings of scalars. *The Quarterly Journal of Experimental Psychology* 71. 272–279. doi:https://doi.org/10.1080/17470218.2017.1314516.

Tessler, Henry M. & Michael Franke. 2018. Not unreasonable: carving vague dimensions with contraries and contradictions. In Charles Kalish, Martina Rau, Jerry Zhu & Timothy T. Rogers (eds.), *Cognitive Science Society*, 1108–1113. Madison, USA.

van Tiel, Bob, Emiel van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33(1). 137–175. doi:https://doi.org/10.1093/jos/ffu017.

Nicole Gotzner
Leibniz-ZAS
Schützenstr. 18
10017 Berlin
gotzner@leibniz-zas.de

Stephanie Solt
Leibniz-ZAS
Schützenstr. 18
10017 Berlin
solt@leibniz-zas.de

Anton Benz
Leibniz-ZAS
Schützenstr. 18
10017 Berlin
benz@leibniz-zas.de

7 Appendix

7.1 Overview of tasks and predictors

Label	Task	Intended measure
New main task SI_mod	Inference judgment (yes\no)	modified scalar implicature
New main task SR	Inference judgment (yes\no)	scale reversal
Main task SI	Inference judgment (yes\no)	scalar implicature
Main task NegS	Inference judgment (yes\no)	negative strengthening
Semantic distance	strength rating (1-7 scale)	scale distinctness
Cloze probability task	free word production	association strength
Politeness weak	kindness rating (1–7 scale)	weak statement
Politeness strong	kindness rating (1–7 scale)	strong statement
Politeness 'not' strong	kindness rating (1–7 scale)	negated strong statement

Table 4 Overview of tasks: two new inference tasks and original results reported in Gotzner et al. (2018).

Label	Predictor	Example
weak min	minimum standard invoked by weaker term	*dirty*
weak max	maximum standard invoked by weaker term	*clean*
weak rel	relative standard invoked by weaker term	*annoyed*
upper bounded	strong term endpoint denoting	*certain*
polarity neg	negative polarity for scale as a whole	*small/tiny*
polarity pos	positive polarity for scale as a whole	*large/gigantic*
relative frequency	log frequency of weak term given strong term	

Table 5 Overview of annotated predictors, details concerning annotation are found in Gotzner et al. (2018).

7.2 Predictor models

	Estimate	SE	t-value	p-value	R^2
(Intercept)	0.25	0.12	2.06		
weak max	-0.11	0.05	-2.18	0.033	
weak min	-0.04	0.03	-1.40	0.166	0.056
upper bounded	0.09	0.03	2.79	0.007	0.133
distance	0.07	0.02	4.11	0.000	0.092
extremeness	-0.15	0.03	-4.54	0.000	0.192
polarity	0.02	0.03	0.88	0.380	0.013
politeness_weak	0.03	0.02	1.18	0.241	0.039
politeness_strong	0.01	0.01	0.51	0.612	0.034
relative frequency	-0.01	0.01	-0.99	0.327	0.022
cloze probability	-0.23	0.15	-1.51	0.137	0.053

Table 6 Predictors of modified scalar implicature task

	Estimate	SE	t-value	p-value	R^2
(Intercept)	0.16	0.14	1.10		
weak max	-0.07	0.06	-1.24	0.219	
weak min	0.04	0.04	1.00	0.323	0.055
upper bounded	0.04	0.04	1.02	0.311	0.016
semantic distance	0.10	0.02	4.89	0.000	0.261
extremeness	-0.08	0.04	-1.98	0.053	0.024
polarity neg	-0.02	0.03	-0.65	0.518	0.003
politeness weak	-0.02	0.02	-0.82	0.413	0.018
politeness strong	0.00	0.02	-0.17	0.862	0.120
relative frequency	0.03	0.01	2.10	0.040	0.092
cloze probability	-0.21	0.18	-1.16	0.251	0.086

Table 7 Predictors of scale reversal task

	Estimate	SE	t-value	p-value	R^2
(Intercept)	1.28	0.32	4.04		
weak min	-0.04	0.04	-0.91	0.37	
weak max	0.15	0.07	2.12	0.038	0.081
upper bounded	-0.07	0.04	-1.64	0.106	0.056
semantic distance	-0.11	0.03	-4.15	0.000	0.184
polarity neg	0.01	0.04	0.32	0.750	0.003
extremeness	0.13	0.04	3.05	0.004	0.085
politeness weak	-0.02	0.02	-0.93	0.357	0.008
politeness not strong	-0.04	0.04	-0.83	0.408	0.011
cloze probability	0.01	0.03	0.37	0.715	0.022
relative frequency	0.26	0.22	1.22	0.228	0.071

Table 8 Predictors of negative strengthening, reprinted from Gotzner, Solt & Benz (2018)

	Estimate	SE	t-value	p-value	R^2
(Intercept)	-0.35	0.25	-1.37		
weak min	-0.01	0.06	-0.22	0.824	
weak max	-0.24	0.10	-2.33	0.023	0.056
upper bounded	0.17	0.07	2.58	0.012	0.114
semantic distance	0.15	0.04	4.03	0.000	0.119
polarity neg	0.02	0.06	0.38	0.703	0.004
extremeness	-0.25	0.07	-3.64	0.001	0.147
politeness weak	0.05	0.04	1.10	0.278	0.025
politeness strong	0.00	0.03	0.03	0.977	0.015
cloze probability	-0.35	0.32	-1.08	0.286	0.053
relative frequency	-0.01	0.03	-0.37	0.712	0.012

Table 9 Predictors of combined measure SI and NegS

	Estimate	SE	t-value	p-value	R^2
(Intercept)	1.16	0.23	5.12		
weak min	-0.03	0.06	-0.49	0.628	
weak max	0.24	0.09	2.60	0.012	0.087
upper bounded	-0.13	0.06	-2.16	0.035	0.059
semantic distance	-0.15	0.03	-4.64	0.000	0.216
polarity negative	0.01	0.05	0.22	0.831	0.001
extremeness	0.16	0.06	2.56	0.013	0.052
politeness weak	0.01	0.04	0.16	0.874	0.002
politeness strong	0.00	0.03	0.15	0.883	0.005
cloze probability	0.37	0.29	1.27	0.208	0.087
relative frequency	-0.02	0.02	-0.99	0.326	0.042

Table 10 Predictors of combined measure NegS and SR

7.3 Second scale reversal task

We also ran a second version of the scale reversal task in which we only presented the weaker term in the conclusion statement. For example, participants judged whether they conclude from *John is not stunning* that *he is brilliant*.

Further, the negative strengthening task was negatively correlated with this second scale reversal task (r_τ = -.34, p <.0001) and there was a positive correlation between the scalar implicature and scale reversal task (modified task: r_τ = .16, p = .05; original task: r_τ = .26, p <.001). Qualitatively, the results were the same in the two versions of the task but correlations were slightly weaker in this second task.

Concerning the predictors, the second scale reversal task yielded different effects in that politeness of the weaker term predicted inference rates which was never a predictor in any previous models. we have to leave it to future research to determine the exact effect of politeness.